Estimating the Null Distribution

James Long
jplong@mdanderson.org
Rice STAT 533 / GSBS 1283

April 21, 2020

Announcements

- HW 9: Due April 23 at 5:00pm, email TA Scott Liang at ricestat533@gmail.com
- Lectures: Today, April 23
- Take home exam
  - Assigned on Thursday April 23
  - Due April 29 at 5:00pm
  - Similar structure to Exams 1 and 2
  - Same policies as Exam 2
  - Strong focus on content in Efron, last third of course
- Lecture Format
  - Slides (plots / analyses in R)
  - .pdf and .R available on course website

Outline

Outline

Two Group Model

- Hypotheses $H_{01}, \ldots, H_{0N}$
- $\pi_0 =$ proportion of true nulls
- $\pi_1 = 1 - \pi_0 =$ proportion of true alternatives
- y_i is indicator H_{1i} is true
  - $y_i \sim Bernoulli(\pi_1)$
- z_i (or p_i) drawn from distribution:

$$f_0(z) \text{ if } y_i = 0 \text{ (i.e. } H_{0i} \text{ is true)}$$
$$f_1(z) \text{ if } y_i = 1 \text{ (i.e. } H_{1i} \text{ is true)}$$

- The marginal distribution of z_i is

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$$

Theoretical Null Distribution

- Null distribution is f_0 (or F_0)
- Theoretical distribution determined by model:
  - $X_{i1}, \ldots, X_{in} \sim_{iid} N(\mu, \sigma^2)$ (the model)
  - $H_{0i} : \mu = 0$
  - $t_i = \bar{X}/(s/\sqrt{n})$
  - Theoretical null distribution for t_i is T_{n-1}
  - $z_i = \Phi^{-1}(F_{T_{n-1}}(t_i))$ where $F_{T_{n-1}}$ is T_{n-1} cdf
  - Theoretical null distribution for z_i is $N(0,1)$ (prefer working with test statistics with $N(0,1)$ null)
- Null distribution necessary for Fdr and local fdr calculations
  - $\overline{\text{Fdr}}(\mathcal{Z}) = \frac{\widehat{\pi}_0 F_0(\mathcal{Z})}{N^{-1}\sum_{i=1}^{N} 1_{z_i \in \mathcal{Z}}}$
  - $\overline{\text{fdr}}(z) = \frac{\widehat{\pi}_0 f_0(z)}{\widehat{f}(z)}$
- Thus far, we have assumed the theoretical null is true
  - **If the model is not a sufficiently good approximation, then the inferences drawn from $\overline{\text{Fdr}}(\mathcal{Z})$ and $\overline{\text{fdr}}(z)$ may be misleading**

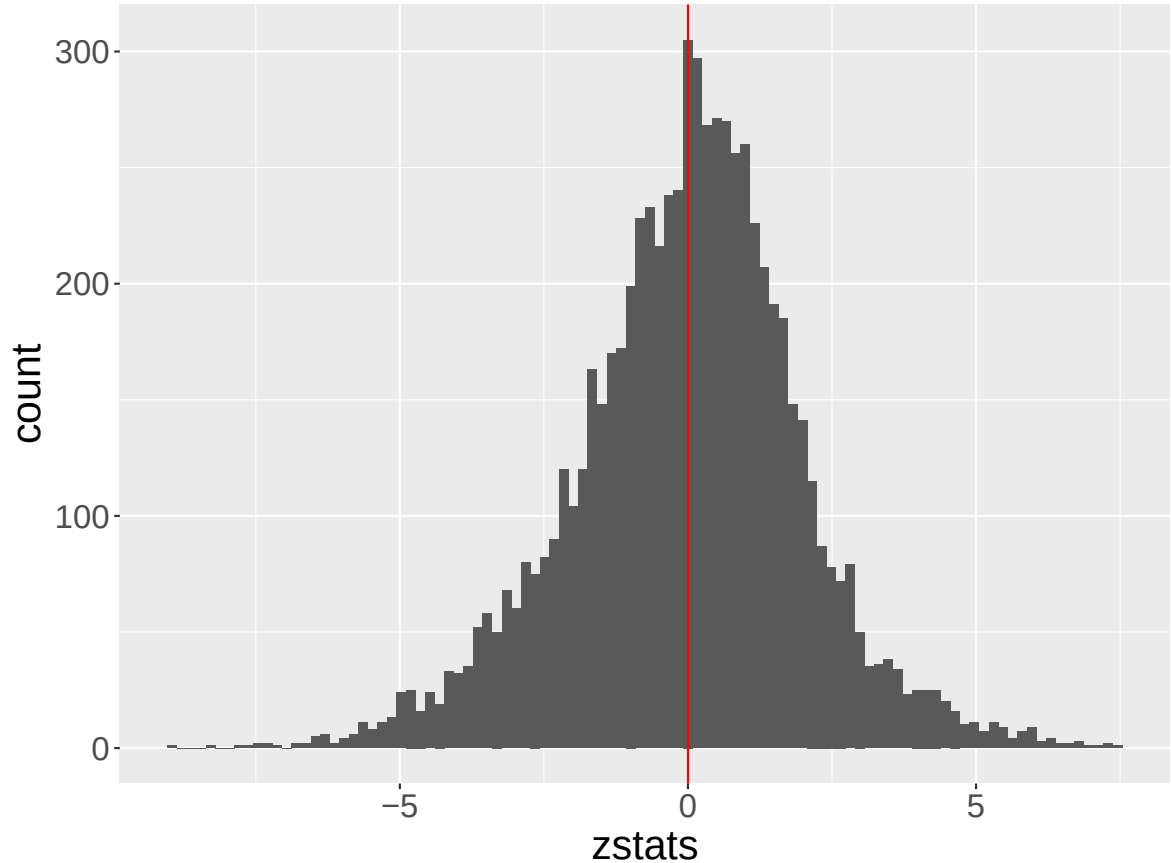
Leukemia Data

- 72 patients with leukemia
  - 47 with ALL (acute lymphoblastic leukemia)
  - 25 with AML (acute myeloid leukemia)
- X_{ij} is expression of gene i for patient j
- Normalize expression values

$$x_{ij} = \Phi^{-1}\left(\frac{\text{rank}(X_{ij}) - 0.5}{N}\right)$$

  where $\text{rank}(X_{ij})$ is rank of X_{ij} among $X_{1j}, \ldots, X_{Nj}$. Removes extreme outliers.
- Two sample t-tests on x_{ij} for $i = 1, \ldots, 7128$

z-statistic Distribution: Theoretical Null is Suspect



- Mean shifted positive
- Highly overdispersed relative to $N(0, 1)$
- Two interpretations:
  - Many true alternatives, i.e. $\pi_0 < 0.75$ (these do not have to follow null)
  - Theoretical $N(0, 1)$ is false

Estimate π_0

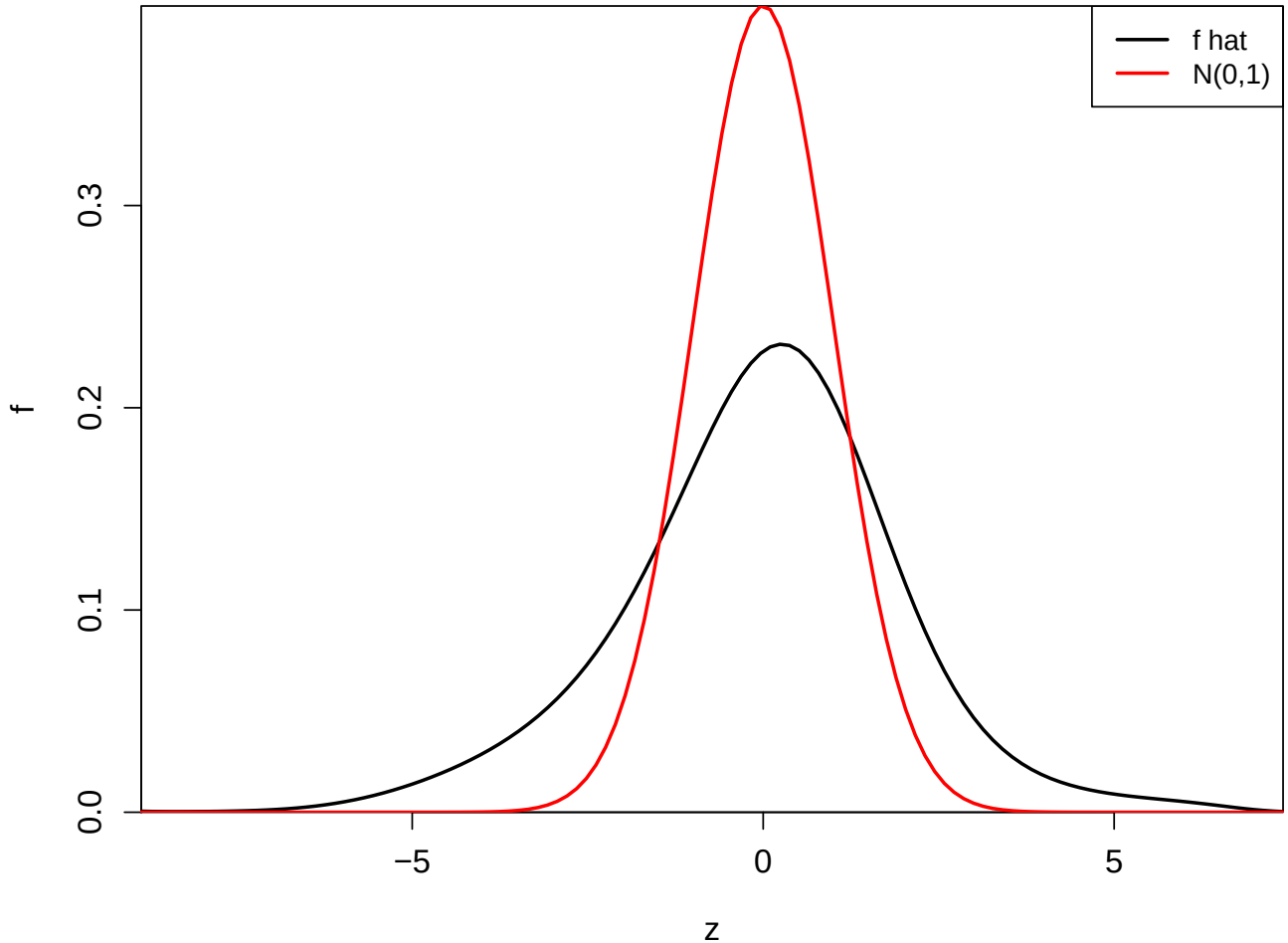
- Estimate π_0 with Null Region Only Method

$$\widehat{\pi}_0 = \frac{N^{-1}\sum_{i=1}^{N} 1_{z_i \in \mathcal{A}_0}}{F_0(\mathcal{A}_0)}$$

- $\mathcal{A}_0 = [\Phi^{-1}(0.25), \Phi^{-1}(0.75)] = [-0.67, 0.67]$
- So

$$\widehat{\pi}_0 = \frac{N^{-1}\sum_{i=1}^{N} 1_{z_i \in \mathcal{A}_0}}{0.5} \approx 0.593$$

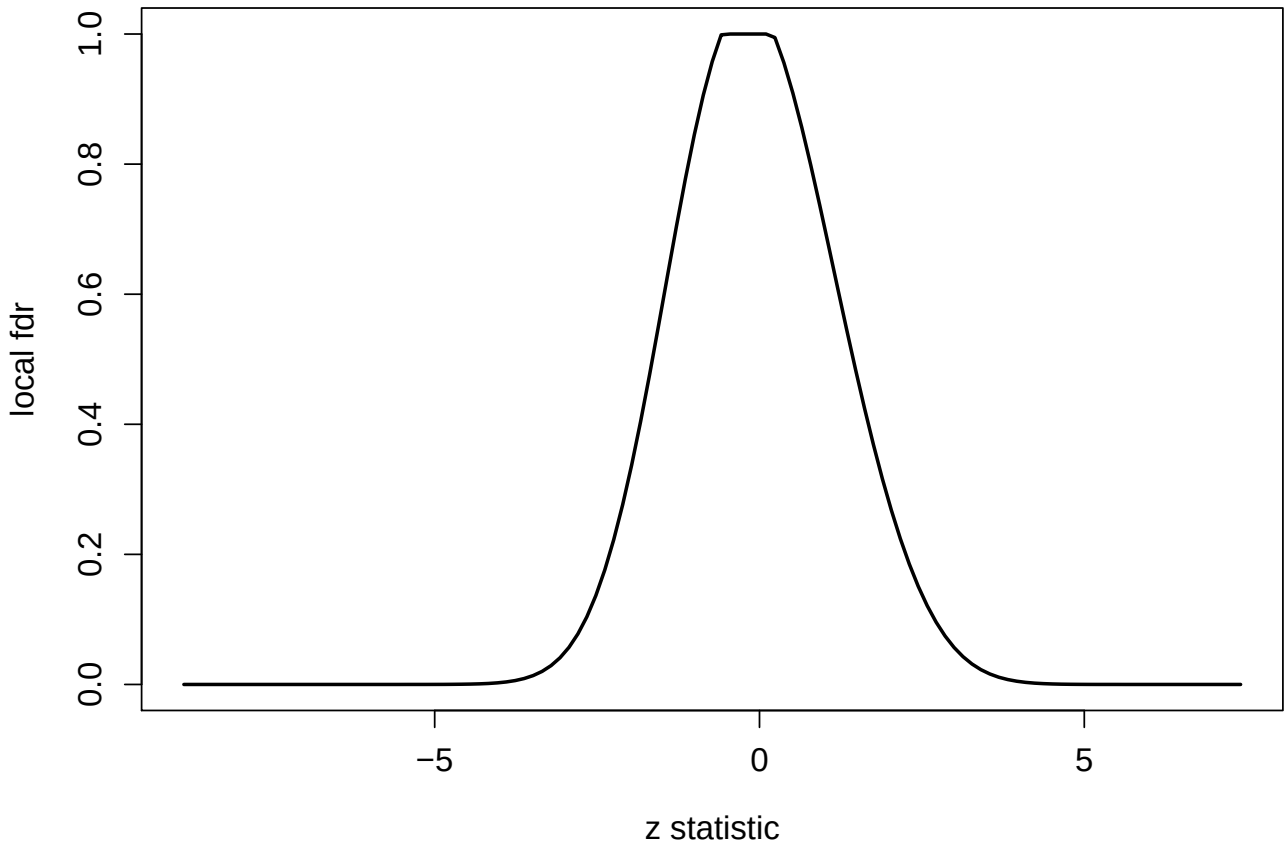
Local fdr



Estimate f with Efron flexible exponential model.

$$\widehat{\text{fdr}}(z) = \frac{\widehat{\pi}_0 \phi(z)}{\widehat{f}(z)}$$

Local fdr



1539 genes with $\widehat{\text{fdr}} < 0.2$

Summary of Results

- $\widehat{\pi}_0 \approx 0.59$ suspiciously low (requires some knowledge about microarrays and cancer to make this determination)
- 1539 genes with $\widehat{\text{fdr}} < 0.2$ is a lot
- Calculations quite sensitive to f_0 (see more later)
- I get somewhat discrepant results from Efron here
  - Efron get $\widehat{\pi}_0 \approx 0.65$
  - Data may be slightly different, sample sizes do not match
  - I use mixture of my own code and Efron's `locfdr` package (on CRAN)
- **Now:** Discuss how / why theoretical f_0 may not be correct null distribution

Outline

Group Distributions

- Leukemia test statistics based on two-sample t-tests
- If AML and ALL not normally distributed, then test statistics may not follow t
  - Quality of normal approximation:
    - If both ALL and AML large, then test statistic is approx $N(0,1)$ regardless of AML/ALL population distributions based on CLT
    - Transform original data to remove outliers
  - Challenges:
    - 25 AML cases, so CLT based normality may not hold
    - Difficult to find transformation that will work for all genes, recall there are 7128 t-tests
- For other models, checking assumptions / improving models even more difficult:
  - Example: Kidney cancer survival Computed test-statistics / p-values using Cox proportional hazards model. Assessing model assumptions / performing transformations with Cox is more difficult than two-sample t-tests

Independence

- Most test statistics assume n independent samples
- For example in two-group t-test x_{ij} are independent across j
- Possible causes of violation in Assumption
  - Matched pairs of patients collected in each group (age, sex, disease progression, etc.)
  - Samples processed in **batches**. For example $x_{i1}, \ldots, x_{ik}$ and $x_{i,k+1}, \ldots, x_{i,n}$ had expression measured on separate trays.
    - Many efforts to remove batch effects, but not always successful.
    - With very poor design, e.g. running all ALL patients in one batch and all AML patients on a second, can completely confound disease and batch effect.

Covariates

- Often collect observational data
- ALL and AML samples not randomized units that were treated with ALL and AML
- Systematic differences in covariates between groups: age, sex, treatments, race, etc.
- Shifts in covariates may cause changes in expression, not related to disease type
- Example:
  - Gene i expression increases with age (all cells)
  - AML patients systematically older than ALL patients
  - Then gene i expression in AML tissue higher than in ALL tissue
  - But differential expression not caused by cancer
  - Healthy tissue from AML patients would have higher gene i expression than healthy tissue from ALL
- **Tricky Issue:** Standard null is wrong for gene i, but not wrong for the reason we care about. Issues related to causal inference.

Fixing the Model

Best solution: Fix model for computing test statistics

- Use test statistics which do not require normality, e.g. Wilcoxon tests instead of t–tests
- Incorporate dependence, e.g. paired t-tests
- Use covariates in model fitting
  - Standard t-test (test statistic computed from α_i):

$$x_{ij} = \mu + \underbrace{\gamma_j}_{j \text{ treat.}=0,1} + \overbrace{\alpha_i}^{\substack{\text{treat.} \\ \text{effect on } i}} + \epsilon_{ij}$$

  - Model with covariates:

$$x_{ij} = \mu + \underbrace{z_j^T}_{\substack{\text{covariates} \\ \text{for } j}} \overbrace{\beta_i}^{\substack{\text{covariate} \\ \text{effect} \\ \text{on gene } i}} + \underbrace{\gamma_j}_{j \text{ treat.}=0,1} + \overbrace{\alpha_i}^{\substack{\text{treat.} \\ \text{effect on } i}} + \epsilon_{ij}$$

Fixing the Model Not Always Feasible

- Only test statistics are available, not gene expression values or covariates
- Computation: More sophisticated models require more computation time
- Division of labor / scientific expertise: Determining appropriate normalization strategies / statistical models for gene expression may involve digging deeply into software pipelines and biological questions about which we have little expertise

Outline

Setup

Assumptions:

1. f_0 follows a (possibly non–standard) normal:

$$f_0(z) \sim N(\delta_0, \sigma_0^2)$$

2. $f_1(z) = 0$ for $z \in \mathcal{A}_0$ (Null only region assumption)

Goal: Estimate π_0, δ_0, σ_0^2 and redo Fdr / local fdr calculations with estimated f_0.

Note: Assumptions are only approximations. May want to consider this as a sensitivity analysis exercise, i.e. How much do results change if consider non $N(0, 1)$ null distribution?

Estimation

Idea:

- Use only z_i in $\mathcal{A}_0$ to estimate δ_0,σ_0^2.
- Use fraction of data in $\mathcal{A}_0$ to estimate π_0

Notation:

- Define

$$r_i = \begin{cases} 1 & z_i \in \mathcal{A}_0 \\ 0 & \text{o.w.} \end{cases}$$

$$z_i' = \begin{cases} z_i & z_i \in \mathcal{A}_0 \\ -\infty & \text{o.w.} \end{cases}$$

- Data is (r_i, z_i') pairs
- $-\infty$ could be any value
- $N_{\mathcal{A}_0} = \sum r_i$ (number of z_i in $\mathcal{A}_0$)
- z_i outside $\mathcal{A}_0$ are "censored"

Likelihood Function

- $p((r_i, z_i')) = p(z_i'|r_i)p(r_i)$
- $p(r_i)$
  - $r_i \sim Bern(\underbrace{\pi_0 H(\delta_0, \sigma_0^2)}_{\equiv\theta})$
  - $H(\delta_0, \sigma_0^2) = \int_{\mathcal{A}_0} \phi(z; \delta_0, \sigma_0^2) dz$
- $p(z_i'|r_i)$
  - Case $r_i = 0$:

$$p(z_i'|r_i = 0) = \begin{cases} 1 & z_i' = -\infty \\ 0 & \text{o.w.} \end{cases}$$

  - Case $r_i = 1$:

$$p(z_i'|r_i = 1) = \frac{\phi(z_i'; \delta_0, \sigma_0^2)}{H(\delta_0, \sigma_0^2)}$$

- Assuming i.i.d. test statistics likelihood is:

$$\prod_{i=1}^{N} p((r_i, z_i')) = \binom{N}{N_{\mathcal{A}_0}} \theta^{N_{\mathcal{A}_0}} (1-\theta)^{N-N_{\mathcal{A}_0}} \prod_{\{i:z_i \in \mathcal{A}_0\}} \frac{\phi(z_i'; \delta_0, \sigma_0^2)}{H(\delta_0, \sigma_0^2)}$$

Notes on Likelihood

- This is censored likelihood common in survival analysis:
  - z_i outside $\mathcal{A}_0$ are treated as censored.
  - Replaced with placeholder $-\infty$
  - r_i is indicator for censoring (1=not censored)
- Log likelihood is concave so MLE is unique
- Could attempt to use original data, e.g. in generalization of BUM / Patra-Sen mixture model:

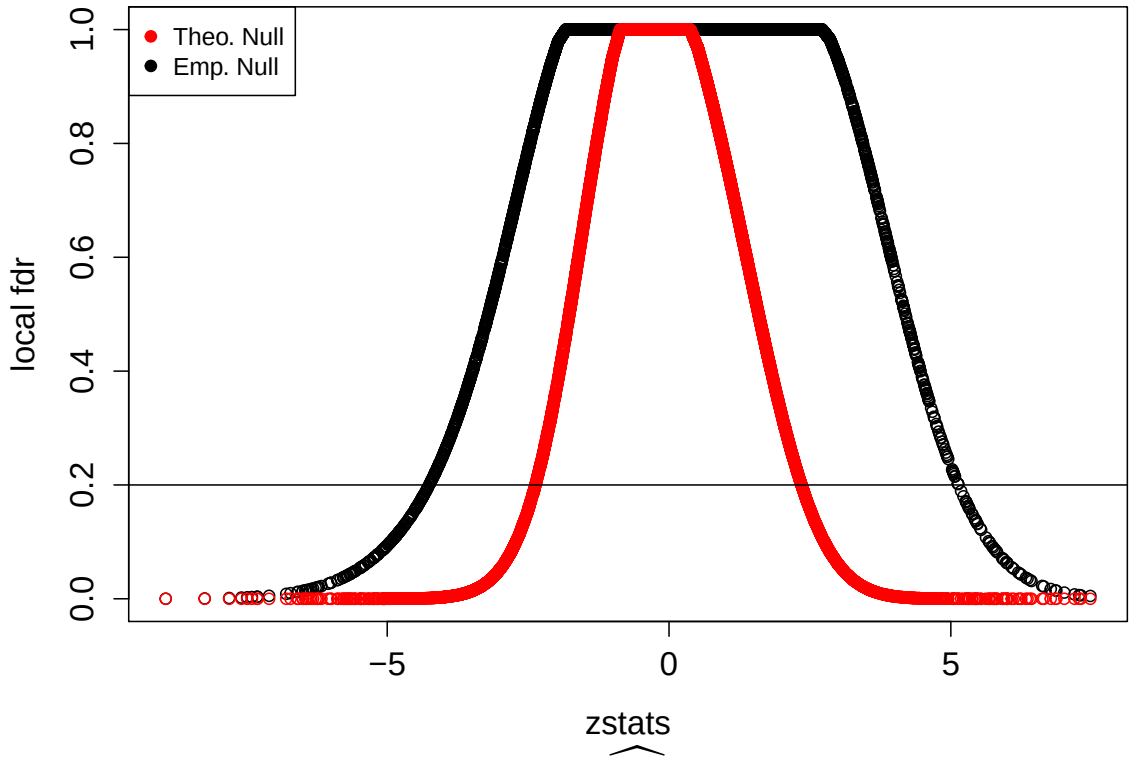
$$F(z) = \pi_0 \Phi\left(\frac{z - \delta_0}{\sigma_0}\right) + (1 - \pi_0) F_1(z)$$

where Φ is $N(0, 1)$ and parameters are σ_0, δ_0, F_1

- Disadvantage: Requires parametric assumptions on F_1 or complex procedures to non-parametrically estimate F_1.

Application to Leukemia Data

Estimates: $\widehat{\delta}_0 = 0.136$, $\widehat{\sigma}_0 = 1.586$, $\widehat{\pi}_0 = 0.915$



Number of hypotheses with local $\widehat{\text{fdr}} < 0.2$:

- Theoretical null: 1501
- Empirical null: 244

Estimating null from data makes a huge difference!

Application to Leukemia Data

- Permutation tests can be used to approximate null distribution
  - Randomly permute AML / ALL labels.
  - Compute test statistics.
  - Use resulting distribution as empirical null.
  - Protects against t-statistic normality assumption.
  - Does not have any effect for Leukemia data (permutation null distribution is approx $N(0, 1)$)
- Leukemia theoretical null violation likely due to hidden covariates.
  - Many genes show small levels of differential expression which are caused by covariates rather than AML/ALL.

Summary / Preview

- Poor theoretical null can greatly change inferences
- Modeling f_0 requires assumptions:
  - f_0 belongs to normal family, Null Only Region near 0
  - Rather than believing the results, can consider modeling f_0 as a type of sensitivity analysis
- Efron suggests that covariates may be biggest problem in causing theoretical null not to hold
- On Thursday Chapter 7: Estimation accuracy of $\widehat{\text{fdr}}$ under dependence